TRANSLATOR FROM SINHALA TO ENGLISH AND ENGLISH TO SINHALA (SEES)

Presented by Laksri Wijerathna
What is this paper about..?

- Rule based machine translation system which is capable of translating sentences from Sinhala to English and vice versa.

- The pilot system this research paper is based on is called “A Sinhala to English and English to Sinhala Translator” - SEES
Problems Addressed

- Language translation has become a necessity for many people as the world moves toward a more global village.

- Majority of people in the globe understand and do their day-to-day work with English.

- When it comes to our country there is a huge barrier between English and Sinhala.

- By developing this pilot system, that barrier can be removed up to some extent.
Challenges

- How to make the machine to do the translation appropriately.
- How to build up a meaningful sentence in the translated set of words.
- How to get the user input sentence and how to make it understandable to the computer.
- What is the atomic level of translation.
How to overcome the Challenges

- Use tokenization to understand the words separately. Use near by tokens to keep track of special cases.

- Use Rule Base technique to make meaningful sentences.

- Get the user input as SinGlish format and store the word and the transaction value in a separate knowledge base.

- Use the sentence as a atomic level translation unit, where every sentence should be ended up with the full stop.
What is SEES

- This system translates a given Sinhala sentence or paragraph into English and vise versa.
Features of SEES

- **Active Voice**
  - Tenses: Present, Past, Future
  - Person Types: 1st Person, 2nd Person, 3rd Person

- **Passive Voice**

- **Sinhala Spell Checker**

- **Sinhala Grammar Checker**

- **Sinhala to English Translator**

- **English to Sinhala Translator**

- **Sinhala Font Translator**
Sinhala to English Translation

Diagram:
- SinGlish Paragraph
  - Unit Identifier
  - Verb Identifier
  - Tokenizer
  - Constructive Translator
  - Verb Translator
  - Sentence Generator
Tokenization

Sudhu
lamaya:
we:gayen:
gedara
dhiw:we:ya.

UNIT

<adj> Sudhu
<sub> lamaya:
<adv> we:gayen:
<noun> gedara
d<verb> dhiw:we:ya.
Verb Stemming

yan:nemu

Different Variations of a Verb

 yan:nemu
 yan:ne:ya
 yan:ni:ya
 yan:no:ya
 yan:nehi
 yan:nehu
### Verb Type and Translation

<table>
<thead>
<tr>
<th>Example</th>
<th>Verb Type</th>
<th>Translation</th>
<th>Additional Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>natamin:</td>
<td>mix verb</td>
<td>dancing</td>
<td>while</td>
</tr>
<tr>
<td>(නටමින්)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kiyawa:</td>
<td>pre verb</td>
<td>read</td>
<td>-</td>
</tr>
<tr>
<td>(කියාව)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pa:dam:</td>
<td>dummy verb</td>
<td>studies</td>
<td>-</td>
</tr>
<tr>
<td>karan:ne:ya</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(කරන:නේ:ය)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sel:lam: karamin:</td>
<td>dummy verb + mix verb</td>
<td>playing</td>
<td>while</td>
</tr>
<tr>
<td>(සැල්මා: කරමින්)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Verb Type – Pre Verb & Mix Verb

- **Sentences with pre verbs**
  - එක සැතපු කියාවේ මම මොගන්න සිහිදුම් කොට.
  - *I read* a book and arranged a short note.

- **Sentences with mix verbs**
  - එක සැතපු කියාවේ සිහිදුම් කොට මොගන්න සිහිදුම් කොට.
  - *She read* a book and gave answers to difficult questions *while laughing*.
Verb Type – Dummy verb

- Sentences with dummy verbs:
  - 证券交易市场.
  - Child studies.

- Sentences with more than one subject:
  - 证券交易市场 证券交易所.
  - Brother tore her book.
English input also tokenized as below.

I went home.

First identify the verb, and then identify the subject and according to the subject verb will be translated.
Verb Translation

Eg: - **went** can be translated as

- **giyemi** (උපෝදේ) - First person, Singular
- **giyemu** (උපෝදෙ) - First person, Plural
- **giya-ya** (උපෝදේ) - Third person, Female, Singular
- **giyo-ya** (උපෝදේ) - Third person, Male, Plural etc.
Noun Translation

- tree - gasa (ගසා)
- a tree - gasak (ගසක)
- from a tree - gasen (ගසේන) etc.

After the translation verb is placed before the full stop.
Research Findings

- Source language provide multiple meaning for the targeted language depending on the word placement.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Output 1</td>
<td>Child <em>crossed</em> the road</td>
</tr>
<tr>
<td>Output 2</td>
<td>Child <em>jumped</em> the road</td>
</tr>
</tbody>
</table>
Source language have the same meaning for multiple words in the targeted language.

<table>
<thead>
<tr>
<th>Input</th>
<th>Output 1</th>
<th>Output 2</th>
</tr>
</thead>
</table>
Ambiguous words of Sinhala Language

<table>
<thead>
<tr>
<th>Input</th>
<th>ohu irak: adin:ne:ya.  අහු පැළකේදාය.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output 1</td>
<td>He draws a line.</td>
</tr>
<tr>
<td>Output 2</td>
<td>He draws a sun.</td>
</tr>
</tbody>
</table>
Future Work

- Increase the Knowledge Base.
- Increase the Rule Base Machine translation system.
- Reduce performance bottlenecks.
- Use a mechanism to identify and handle the ambiguous words and do the translation accurately.
- Add self learning feature to the system.
Thank You...!